

HIV Database Workshop

www.hiv.lanl.gov
seq-info@lanl.gov

Presenter: Bette Korber

Database staff: Werner Abfalterer, Chuck Calef, Robert Funkhouser, Shaun Geer, Brian Gaschen, Kristina Kommander, Bette Korber, Dorothy Lang, Thomas Leitner, James Szinger, Karina Yusim, Ming Zhang

Regular Contributors: John Mellors and Christian Brander

Project Officer: James Bradac, NIAID, NIH

HCV Databases: Carla Kuiken and Karina Yusim: www.hcv.lanl.gov



*Theoretical Biology and Biophysics, T-10
Los Alamos National Laboratory*

HIV Database workshop



Workshop Topics

June 9th, 2005, Durban South Africa

Introduction

Sequence Database

Search tools *Basic sequence search interface and on-the-fly alignments and trees*
HIV/SIV sequence locator tool
SynchAligns and Treemaker
Geography search interface

New Sequences *Genecutter - processing nucleotide sequences*
HIV database alignments and subtype reference sequences

Analysis Tools *Using the new RIP tool for recombination analysis*
Contamination
Nglyco

Immunology Database

Database *CTL search page*
Ab search page
Epitope maps

Tools *Epiline*
Motif Scan
Hepitope
ELF

About the Instructor

Dr. Bette Korber is Co-PI of the database project, primary editor of the HIV Immunology Database, with background in HIV evolution and immunology



Workshop Goals

- Understanding the database content, how information was obtained, and what is available
- Quality control tools
- Tools for analyses
- Database searching

The HIV Databases

- HIV Sequence database – founded 1986, G. Myers
 - Relational database, data from GenBank with added fields from the literature
 - Alignments – align indels and reduce multiple sequences per person
 - Annual hard copy and reviews
 - Web search interfaces: subtype, phenotype, geographic, sampling year...
 - Analysis tools
- HIV Immunology database – founded 1995, B. Korber
 - Comprehensive HIV epitope database, 300-400 papers a year
 - Integrate HIV immunological and sequence data
 - Annual hard copy and reviews
 - Web search interfaces: epitope, protein, HLA type, immunogen, keywords
 - Analysis tools for immunologists
- HIV Drug Resistance database, founded 1997, J. Mellors
 - A searchable web listing of drug resistance mutations and literature links, updated annually by Dr. Mellors
- HIV Vaccine database, founded 2003, J. Mokili
 - A searchable relational database of published primate vaccine trials



 **HIV Databases** 

The HIV databases contain data on HIV genetic sequences, immunological epitopes, drug resistance-associated mutations, and vaccine trials. The website also gives access to a large number of tools that can be used to analyze these data. This project is funded by the Division of AIDS of the National Institute of Allergy and Infectious Diseases (NIAID), a part of the National Institutes of Health (NIH). Click on any of the links below to access a database.

[Sequence Database](#)
[Resistance Database](#)
[Immunology Database](#)
[Vaccine Trials Database](#)

[HCV Databases](#)

news

- To avoid future confusion in CRF and subtype designations, the HIV nomenclature committee has agreed that for us to assign new CRF numbers or subtype letters we must be given the sequence and mosaic pattern maps of potentially new CRFs and subtypes before we can assign new names. See further [The Circulating Recombinant Forms](#). 25 January 2005
- The 2003 HIV and SIV Sequence Alignments are now available [online](#). 25 December 2004
- The HIV Sequence Compendium 2003 is now available [on-line](#). The printed copy is also being printed and shipped now. 17 December 2004

[Old News](#)

Questions or comments? Contact us at seq-info@t10.lanl.gov

HIV sequence database main page - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Home Search Favorites Media Print W

Address <http://hiv-web.lanl.gov/content/hiv-db/mainpage.html> Go Links



Search Site

Databases

- Sequence DB
- Resistance DB
- Immunology DB
- Vaccine trials DB

Publications

- FAQ
- Alignments
- Tutorials
- Reviews
- Compendia
- Links

Sequence DB

- Search DB
- Tools
- HIV-Blast
- Recombination
- Syn-Nonsyn
- Hypermut
- PCOORD
- SUDI
- Treemaker
- Geography
- N-Glycosite
- 3D Structure
- GeneCutter
- RIP 2.0
- External Tools

HCV Databases

- SynchAligns
- Findmodel

News [Old News](#)

- New sequence database [search fields](#) are available: viral load, CD4 and CD8 counts, ethnicity, and cohort. These field are poorly filled out at the moment, and this information is not available for most sequences, but over the next few years we will be striving to get this information whenever possible, particularly emphasizing large cohorts. A large data set from Durban South Africa is particularly well represented. 26 April 2005
- A new screen enabling users to exclude [problematic sequences](#) has been added to the HIV sequence search interface. 30 March 2005

About this website:

- [Overview of the site](#)
- [Frequently asked questions](#)

Programs and tools:

- New! Search interface (Beta):** Adds the ability to create neighbor-joining trees from the query results.
- Search interface:** retrieve sequences based on all HIV database search fields. HIV-1 sequences can be aligned and clipped. This interface combines the HIVMAP and DBSearch interfaces. On the search interface webpage, a link to the old search interface is still available.
- Tools for working with sequences:** GeneCutter, SeqConvert, Gapstrip, Motifscan, Primalign, Epilign, HXB2 Numbering, SeqPublish, HIV-BLAST, sequence format conversion, N-GlycoSite, Entropy
- Programs for sequence analysis:** RIP (Recombination), SNAP (Syn-Nonsyn), VESPA (Signature patterns), Hypermut, PCOORD, and more...
- Other programs** for analysis of HMA data and optical density data
- Links to external programs:** phylogenetics, recombination, subtyping, multiple alignment, sequence submission

Tutorials:

- [Sequence quality control](#) January 1998
- [How to build a phylogenetic tree](#) October 1999
- [HIV and SIV subtype nomenclature](#) September 2000
- [How to use these databases -- workshop given at 11th Conference on Retroviruses and Opportunistic Infections](#) February 2004

Alignments:

- [Complete alignments](#) of all genes (nt and aa) and complete genomes (nt only)
- [Subtype reference alignments](#) for use in trees, subtype comparisons and recombination research
- [Consensus and Ancestral sequences](#) for M_GROUP and subtypes

Compendia:

- [Print \(PDF\) or order a copy](#) of our compendia "Human retroviruses and AIDS"
- [Reviews published in the Sequence and Immunology Compendia](#)
- [How to refer to the compendia](#) in a publication

Links:

- [Los Alamos Immunology Website](#), our sister site, houses a huge searchable collection of HIV immunological epitopes
- [Los Alamos Drug Resistance Database](#) contains information about anti-HIV drugs and drug-resistance-conferring mutations
- [Los Alamos HIV/SIV Vaccine Trial Database](#) A database of vaccine trials, including design, type of vaccine used, results, etc.
- [Other HIV/AIDS sites](#) for more software and information on HIV/AIDS

<http://hiv-web.lanl.gov/HTML/FAQ.html> Internet

Search Interface

Help

- ☐ Tips at the top of the page are often overlooked
 - Ranges, operators, wildcards, logical groupings
- ☐ Field names are clickable, also mouse-overs
 - Example: "Sampling country" gives two-letter ISO country codes

Searches

- ☐ Searches are case-insensitive
- ☐ Records are searchable through sequence, patient, genomic region, or publication information
- ☐ First seven fields will appear in search results page by default
- ☐ A "*" in a textbox will cause that field to be included in the results page
- ☐ Patient information (Infection year, Infection country) is different than sequence information (Sampling year and Sampling country)

Results

- ☐ Can select not aligned, or aligned based on multiple pair wise alignments – alignments are good, but still need hand editing for an optimal alignment
- ☐ Select all or a subset of sequences for download
- ☐ Sequences can be re-ordered by clicking on fields at the top of the page

HIV Sequence Search Interface - Microsoft Internet Explorer

Address: http://hiv-web.lanl.gov/components/hiv-db/combined_search_s_tree/search.html

Advanced Search

Subtype: Any

Organism: Any

Incl problematic seqs: ☐ yes ☒ no

Infection country:

Infection year:

CD8 count:

Phenotype: list field in output

Incl only drug naive sequences: ☐ yes ☒ no

Months from seroconversion:

Months post infection:

Other fields:

Use this option to search based on the HIV-DB internal alignment (presently this option restricts the search to HIV-1):

☐ Include fragments of minimum length 100

Genomic region: Any

Or define start: and end:

Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print

Address http://hiv-dev.lanl.gov/components/hiv-db/combined_search_s/search.comp

Please note that if a genomic region is specified, only HIV-1 sequences can be found

Search Site

Databases

- Sequence DB
- Resistance DB
- Immunology DB
- Vaccine trials DB

Publications

- FAQ
- Alignments
- Tutorials
- Reviews
- Compendia
- Links

Sequence DB

- Search DB
- Tools
- HIV-Blast
- Recombination
- Syn-Nonsyn
- Hypermut
- PCOORD
- SUDI
- Treemaker
- Geography
- N-Glycosite
- 3D Structure
- GeneCutter
- RIP 2.0
- External Tools

HCV Databases

- Internal
- Style Help

Disclaimer/Privacy

Build a Tree or Download Sequences

Use Sequences

☒ Clip to selected region

☐ Include HXB2 Reference Sequence (K03455)

Show names as

Displaying 1 - 6 of 6 sequences found:

[Select all](#) [Unselect all](#) [Invert selection](#)

[Select](#) record to [List](#) records per page

Click on field name to sort in ascending or descending order

#	Select	Patient Code (id)	Accession	Name	Subtype	Country	Sampling Year	Infection Country	Genomic Region	Sequence Length	Organism
1	<input type="checkbox"/>	Blast NH1(57553)	AB052995	93JP_NH1	01_AE	JP	1993	TH		9720	HIV-1
2	<input type="checkbox"/>	Blast NH2(57551)	AB070352	NH25 93JPNH25T 93JP_NH2_5T	01_AE	JP	1993	JP		9731	HIV-1
3	<input type="checkbox"/>	Blast NH2(57551)	AB070353	NH2 93JPNH2ENV	01_AE	JP	1993	JP		9720	HIV-1
4	<input type="checkbox"/>	Blast NH1(57553)	BD187399	93JP_NH1	01_AE	JP	1993	TH		9720	HIV-1
5	<input type="checkbox"/>	Blast ETR(10141952)	D12582	ETR	B	JP				2589	HIV-1
6	<input type="checkbox"/>	Blast JH32(6062)	M21138	JH32	B	JP	1986			2903	HIV-1

[Download](#) tab-delimited results, include sequence ☐

Last modified: Thu Feb 17 14:49 2005

Questions or comments? Contact us at seq-info@t10.lanl.gov

Search Results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print

Address http://hiv-dev.lanl.gov/components/hiv-db/combined_search_s/search.comp

Search Site

Databases

- Sequence DB
- Resistance DB
- Immunology DB
- Vaccine trials DB

Publications

- FAQ
- Alignments
- Tutorials
- Reviews
- Compendia
- Links

Sequence DB

- Search DB
- Tools
- HIV-Blast
- Recombination
- Syn-Nonsyn
- Hypermut
- PCOORD
- SUDI
- Treemaker
- Geography
- N-Glycosite
- 3D Structure
- GeneCutter
- RIP 2.0
- External Tools

HCV Databases

- Internal
- Style Help

Disclaimer/Privacy

Tree Builder

This tool produces trees with a neighbor-joining algorithm.
Note: These "quick and dirty" trees are not intended for use in publications.
Learn how to make a Phylogenetic Tree

[go](#)

Subtype Reference Sequences to include:

None
All subtype reference sequences
All M Group (A-K) subtype reference sequences
All Non-Recombinant subtype reference sequences
A1 KE 93.Q23-17 AF004885
A1 SE 94.SE7253 AF069670
A1 UG 85.U455 M62320
A1 UG 92.92UG037 U51190
A2 CD -.97CDKTB48 AF286238
A2.CY.94.94CY017.41 AF286237

DNADist Parameters:

Distance Model

Transition/Transversion Ratio (ignored for Jukes-Cantor model)

Neighbor Parameters:

Show names as

Outgroup:

Selected sequence outgroup OR... Subtype reference sequence outgroup

[reset](#) [go](#)

This tree-building tool uses [PhyIip](#)

Last modified: Wed Feb 16 16:40 2005

Questions or comments? Contact us at seq-info@t10.lanl.gov

HIV/SIV Sequence Locator Tool

- Rapidly returns position numbers of an HIV or SIV DNA or protein sequence fragment relative to the HXB2r or SMM239 reference strains
- Such numbers are often included in the literature, and are often incorrect
- Marks the location of the sequence on an HIV map
- For DNA sequences, a translation is provided
- Can be used for input into the search interface, to pull a particular region of interest out of the database (like a specific epitope)

HIV/SIV Sequence Locator Tool

Link to old HXB2 Numbering Engine | Numbering Positions in HIV Relative to HXB2

Paste your sequence in the box below, or use the browse button to select a file that contains sequence(s) to upload. You can specify that your input sequence(s) are HIV or SIV by selecting the appropriate choice from the list on the right, or you can let the program decide (default). [Details](#).

If you submit multiple sequences then indicate that here -->

☐ Examine reverse complement of sequence

Input Options: Default is "Single sequence". See options in table below. You can mix nucleotide and amino acid sequences in your input.

Single sequence	Multiple sequences
<p>1. Free format. Sequence can contain carriage returns, spaces, dashes, and other characters. The program will remove all non-letter characters like returns and spaces before processing.</p> <p>Example input:</p> <pre>act gatgc---tcagtcg xactt agn tagtcga</pre> <p>will be treated as</p> <pre>actgatgctcacgtatcgxacttagntagtcga</pre>	<p>To submit multiple sequence the user must select the "Multiple Sequences" option. The program recognizes two multiple sequence input options:</p> <p>1. Fasta format. Example:</p> <pre>>seq1 LAREEVVIRSENFDTNAKTIIVQLN ESVRINCTRPNNN >seq2 GPCRAFYTTGQIIIGDIPQAH >seq3 VTKLREQFKN-KTIVFNQSSGCD</pre> <p>The program recognizes this format by the ">" character. All non-letter characters will be removed, so it is OK if your sequence contains returns (e.g. seq1) or gaps (seq3).</p> <p>2. Raw sequence, one sequence per line with a carriage return between sequences. Example:</p> <pre>LAREEVVIRSENFDTNAKTIIVQLNESVEINCTRPNNN GPCRAFYTTGQIIIGDIPQAH VTKLREQFKN-KTIVFNQSSGCD</pre>

HIV or SIV? You can specify that your input sequence(s) are HIV or SIV by selecting the appropriate choice in the list, or you can let the program decide.

Comparing "new" and DB sequences

- Synch Aligns allows you to bring your new alignment into register with one of our reference alignments or search alignments
- TreeMaker produces a Neighbor Joining tree for a "quick-and-dirty" comparison
- TreeMaker is based on DNADIST & NEIGHBOR in the PHYLIP package
- HIV-BLAST is an option for looking for highly similar sequences or possible contamination

SynchAligns - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://hiv-web.lanl.gov/content/hiv-db/SYNCH_ALIGNS/SynchAligns.html Go Links

Synchronize Alignments

Purpose: This tool aligns two alignments to each other. It does this by using a sequence, called the reference sequence, that is common to both alignments. If the two alignments do not share a common sequence, the program chooses the longest sequence from each alignment as the references, aligns them to one another and then adjusts the two submitted alignments to agree with the aligned references. Here's an example.

Explanation: *Input format:* The program should correctly read any valid format. Each sequence must have a name, so you cannot submit raw sequence files. The files can be in different formats, but the output format will be the same as the input format of the second file. Also specify the gap character if it is not a dash (-). More than one gap character may be specified in case your two alignments use different characters.

Reference sequences: The program will synchronize your files using two reference sequences that it selects. If this fails and your alignments share a common reference sequence, you can check the "Reference sequence selection: Manual" box; then you will be asked to identify this sequence in both alignments. In this case, the program does not require that the reference sequences have the same names, but the sequences must be identical.

Please note: This program can make changes to the order of the sequences; lower case sequences get changed to upper case; all gap characters are changed to '-'; and uneven seqs get padded with '-' to make them the same length.

How to use: Upload your two alignment files.

Alignment 1 Browse...

Alignment 2 Browse...

Gap characters:

Squeeze gaps from input ☒

Reference sequence selection ☒ Automatic ☐ Manual

Trim alignments to region of overlap ☐

Last modified: Tue Mar 29 15:00 2005

Questions or comments? Contact us at seq-info@t10.lanl.gov

Synchronize Alignments Example

```
ref1      JKLMN-OPQR-ST
align1    JKLMNYOPQRYST
          JKLMNYOPQR-ST

ref2      HIJK-LMNOP
align2    HIJKXLMN-P
          --JK-LMNOP
          HIJK-LMNOP

Result after SynchAligns

ref1      --JK-LMN-OPQR-ST
align1    --JK-LMNYOPQR-ST
          --JK-LMNYOPQRYST


ref2      HIJK-LMN-OP-----
align2    HIJK-LMN-OP-----
          H-JK-LMN-OP-----
          HIJKXLMN-OP-----
```

HIV Sequence Database Treemaker - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Search Favorites Media Print View Source

Address <http://www.hiv.lanl.gov/content/hiv-db/CONTAM/TreeMaker/TreeMaker.html> Go Links



Search Site

Databases

Sequence DB
Resistance DB
Immunology DB
Vaccine trials DB

Publications

FAQ
Alignments
Tutorials
Reviews
Compendia
Links

Sequence DB

Search DB
Tools
HIV-Blast
Recombination
Syn-Nonsyn
Hypermut
PCOORD
SUDI
Treemaker
Geography
N-Glycosite
3D Structure
GeneCutter
RIP 2.0
External Tools

HCV Databases

Disclaimer/Privacy

NEIGHBOR TREEMAKER

Neighbor TreeMaker takes a sequence alignment, converts it to [PHYLIP](#) format, runs it through the PHYLIP programs **Dnadist** (Distance Matrix program), and **Neighbor** (treefile generator), then displays a tree.

The **Dnadist** program reads in nucleotide sequences and writes an output file containing the distance matrix. PHYLIP gives options concerning the model used to calculate distances. This interface uses ML, the model used in PHYLIP's maximum likelihood phylogeny program DNAML. This model incorporates different rates of transition and transversion, and also allows for different frequencies of the four nucleotides. [PHYLIP copyright/reference info](#).

***Disclaimer.** This interface only offers very basic, 'quick-and-dirty' phylogenetic analysis. More in-depth analysis is usually needed. For more information see the [Treemaker Tutorial](#).*

1. Submit Alignment.

Please paste your alignment into the submission box below and **indicate the format**. This interface only accepts nucleotide sequences.

Paste your alignment here:

Current Format: FASTA Sample Input

Browse...

2. Set Program Parameters.

You can tune **Neighbor TreeMaker** by adjusting two parameters.

a. specify the number of the **outgroup** sequence in your alignment (to be the root of your tree; default is the first sequence):

b. specify the **transition/transversion ratio** 1.30

3. Choose Program Output.

The tree can be presented as a

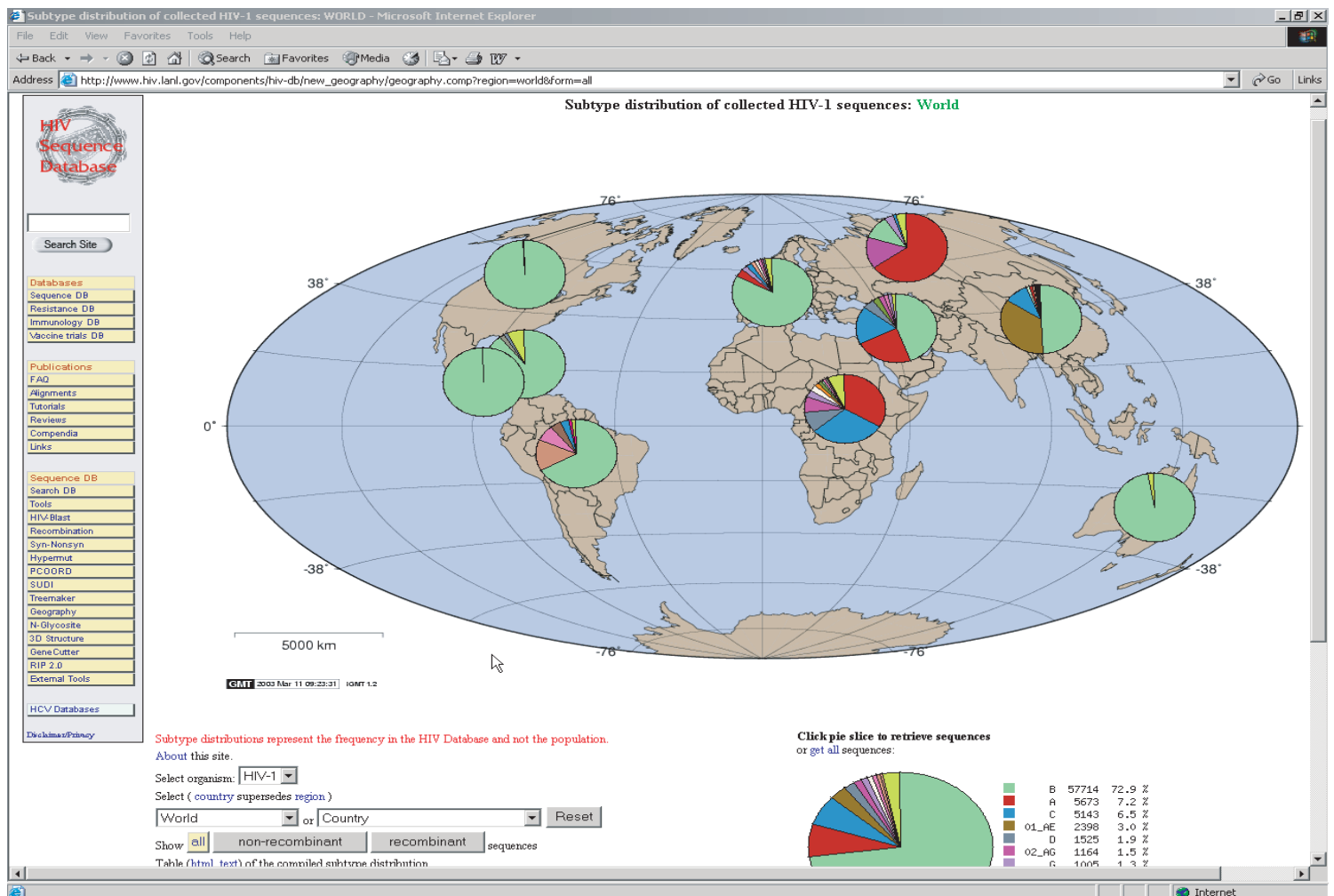
☒ Phenogram or as a

☐ Radial Tree

(Once the tree has been generated, you can view the text treefile and outfile.)

Geography Tool

- Another way to search/download sequences is by geographic region or country
- Results are biased as they show only the sampled individuals, not the true subtype distribution for a region's population
- Results are selectable as in the search interface



Gene Cutter

- Useful for sequencing labs, particularly for rapid processing of new sets of full length genomes
- Cut out genes and proteins from aligned sets of DNA sequences
- Sequences do not need to be codon aligned – results can be codon-aligned on the fly with generally good results
- Currently, sequence alignments must contain HXB2 as a reference for the program to function
- **New Features**
 - Allows codon alignment and cutting of regions of HIV-2 and SIV (must include SMM239)
 - Recognizes IUPAC multistate characters in sequences and translates them appropriately
 - For each region, maintains a list of stop codons, codons containing multistate characters, and codons containing indels.

GENE CUTTER

Gene Cutter is a tool that clips pre-defined coding regions from a nucleotide alignment, then codon aligns and provides translations of the cut regions.

GeneCutter requires the proper reference sequence (HXB2 - K03455 for HIV-1 alignments) (SMM239 - M33262 for HIV-2 or SIV alignments) to be included in your input nucleotide alignment. It uses the reference sequence to set the proper reading frame for the rest of the alignment by slightly rearranging the sequence alignment so that the reference is in frame.

Remember: Your output alignment will only be as good as your input alignment.

Note: In some sequences an insertion will be compensated for within a short distance by a deletion, or vice versa. As these frameshifts may not inactivate the protein, if a compensating mutation is within 5 amino acids of an initial frameshift, the frame-shifted reading frame is left intact. Otherwise, the frame shift is marked with the hash symbol (#), and the translation is continued in the correct, typical reading frame beyond the offending codon. Stop codons are marked by a dollar sign (\$).

Input options:
Select the region of the alignment you would like to extract. Note: The reference sequence contained in your alignment **does not** need to cover all of the selected region to be cut correctly.

HIV1 (HXB2) [v]
All proteins [v]

Enter your **alignment containing either HXB2 (HIV-1) or SMM239 (HIV-2/SIV)** here.

Sample Input

[Text input field]

[Browse...]

Input alignment format: [FASTA v]

Output Options:

Sequence Alignments

- Originally based on iterations of manual and HMM alignments
- Yearly updates using HMM and manual corrections
- Full length genomes updated throughout the year
- Alignments are in reading frame (codon aligned)
- Alignments non-redundant
- Compendia alignments show fewer sequences than web version
- Reference alignments contain up to four representatives
- Protein alignments may contain frameshift compensations
- Subtype consensus with ties resolved, as well as maximum likelihood ancestors, are available for reagent production

The screenshot shows the HIV Sequence Database website in a Microsoft Internet Explorer browser window. The address bar shows the URL: http://www.hiv.lanl.gov/content/hiv-db/ALIGN_CURRENT/ALIGN-INDEX.html. The page title is "2003 HIV and SIV alignments".

On the left side, there is a navigation menu with the following sections:

- Databases**
 - Sequence DB
 - Resistance DB
 - Immunology DB
 - Vaccine trials DB
- Publications**
 - FAQ
 - Alignments
 - Tutorials
 - Reviews
 - Compendia
 - Links
- Sequence DB**
 - Search DB
 - Tools
 - HIV-Blast
 - Recombination
 - Syn-Nonsyn
 - Hypermot
 - PCOORD
 - SUDI
 - Treemaker
 - Geography
 - N-Glycosite
 - 3D Structure
 - GeneCutter
 - RIP 2.0
 - External Tools
 - HCV Databases
- Disclaimer/Privacy

The main content area has a heading "2003 HIV and SIV alignments" and a note: "Note : The protein alignments provided for each gene were constructed using both nucleotide and translated amino acid sequences. Because the translations are based on alignments, they may differ from a straight, non-aligned, translation. For instance, an aligned translation will include frameshift compensation."

Below the note, it states: "These alignments were generated by an iterative process between automated alignment using HMMER and manual editing using MASE, BioEdit and Se-Al. Any alignment presented is not suggested to be an 'optimal alignment' with the absolute minimum number of gaps and mismatches. It is a compromise between optimal alignment, readability, and an attempt to keep insertions and deletions from altering the protein reading frame presentation. Especially the 'Other SIV' alignments are difficult to make, consider these as a starting point for your analyses. Most gaps have been introduced in multiples of 3 bases to maintain open reading frames when translated directly from the alignment."

Below this, it says: "Codons containing IUPAC/TUB multistate characters involved in silent substitutions are translated to amino acids, otherwise they are translated to 'X'."

There is a dropdown menu for "Alignment format:" set to "FASTA". Below it, a link says: "Click here for information on color-coded protein alignments."

Below the link, there are two buttons: "Get Alignment" and "reset".

Below the buttons is a table with the following structure:

Region	HIV-1/SIVepz	HIV-2/SIVsmm	Other SIV
Genome	<input type="radio"/> DNA	<input type="radio"/> DNA	<input type="radio"/> DNA (includes HIV-1 and HIV-2 sequences)
LTR	<input type="radio"/> DNA	<input type="radio"/> DNA	<input type="radio"/> DNA
GAG	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
POL	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
ENV	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
VIF	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
TAT	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
REV	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
VPV/VPX	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
VPR	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein
NEF	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein	<input type="radio"/> DNA <input type="radio"/> Protein

At the bottom right, it says "Last modified: Sat Dec 25 00:00 2004".

At the bottom center, it says "Questions or comments? Contact us at seq-info@t10.lanl.gov".

At the bottom left, there is a link: <mailto:seq-info@t10.lanl.gov>.

At the bottom right, there is a link: [Internet](#).

Recombination Analysis

- Many methods and programs exist to investigate potential recombination
 - <http://bioinf.man.ac.uk/~robertson/recombination/>
- Investigating recombination requires many steps
- A new version of RIP is available at HIV db
 - Automatic alignment
 - Selection of background sequences
 - Different window/gap handling options
 - Graphic & table output

The screenshot shows the 'RIP 2.0' submission page of the HIV Sequence Database. The page is divided into three main sections: 'QUERY SEQUENCE', 'BACKGROUND ALIGNMENT', and 'RIP SETTINGS'. On the left, there is a sidebar with navigation links for 'Databases', 'Publications', 'Sequence DB', and 'HCV Databases'. The 'QUERY SEQUENCE' section has a 'Format of your query sequence' dropdown set to 'FASTA', an 'Upload your query sequence file' button, and a 'Sample Input' button. Below this is a text area for pasting the query sequence. The 'BACKGROUND ALIGNMENT' section has radio buttons for 'Use subtype consensus sequences as background' (selected), 'Include 01_AE consensus', 'Exclude 01_AE consensus', 'Select custom background from subtype consensus and representative sequences', and 'Use your own alignment: query (first sequence) aligned to background'. It also has an 'Upload your alignment' button and a text area for pasting the background alignment. The 'RIP SETTINGS' section has a 'Window size' dropdown set to '200', a 'Significance threshold' dropdown set to '90%', and 'Gap handling' radio buttons for four options: '1. Treat gaps as characters; Plot all window values', '2. Treat gaps as characters; Don't plot window values for gaps in query.', '3. Strip all gaps; Plot all window values', and '4. Strip all gaps; Plot blanks where gaps used to be.'. The 'Output format' section has radio buttons for 'Graphical', 'Tabular', and 'Both'.

RIP 2.0

This is the sequence submission page for the new version of RIP. In this version of RIP your query sequence is automatically aligned to the background alignment. The alignment so produced may be downloaded from the results page. The output of this program should be interpreted with caution. Please [send us](#) bug reports or suggestions.

QUERY SEQUENCE

Format of your query sequence:

Upload your query sequence file ...

or paste your query sequence here:

BACKGROUND ALIGNMENT

☒ Use subtype consensus sequences as background

☒ Include 01_AE consensus ☐ Exclude 01_AE consensus

☐ Select custom background from subtype consensus and representative sequences

☐ Use your own alignment: query (first sequence) aligned to background

Upload your alignment ...

or paste your background here ...

RIP SETTINGS

Window size

Significance threshold

Gap handling

☒ 1. Treat gaps as characters; Plot all window values

☐ 2. Treat gaps as characters; Don't plot window values for gaps in query.


☐ 3. Strip all gaps; Plot all window values

☐ 4. Strip all gaps; Plot blanks where gaps used to be.

Output format ☐ Graphical ☐ Tabular ☒ Both

Check new sequences to make sure they are “sensible”

- Trees – are intrapatient and linked cases well behaved?
- Blast – use representative sequences screen against the database to check for possible contaminations.
- Check alignments – look for regional odd behavior – PCR recombination can carry in contaminating fragments
(coming soon: window blast)
- Are sequences hypermutated?



Search Site

Databases

- Sequence DB
- Resistance DB
- Immunology DB
- Vaccine trials DB

Publications

- FAQ
- Alignments
- Tutorials
- Reviews
- Compendia
- Links

Sequence DB

- Search DB
- Tools
- HIV-Blast
- Recombination
- Syn-Nonsyn
- Hypermut
- PCOORD
- SUDI
- TreeMaker
- Geography
- N-Glycosite
- 3D Structure
- GeneCutter
- RIP 2.0
- External Tools

HCV Databases

Disclaimer/Privacy

Sequence Quality Control

"No matter what drug we give, our sequences always are 25% wildtype"
- Quote from the head of a reference laboratory.

Some HIV researchers are convinced that careful lab work is enough to prevent contamination. We disagree. Contamination happens, even in the best laboratories. Screening for contamination should be done before the analysis of the sequences, and periodically during the course of large sequencing studies, so problems can be detected and corrected early.

To show what contamination looks like in practice, we have collected some [examples](#) of (mostly published) datasets where contamination is a problem, and included some references that discuss contamination.

Following the steps below will help to check your sequences. They are no substitute for common sense and precautions, but they may help spot contamination in your sequences. We have created interactive pages where you can build a tree and do a BLAST search with your sequences. If you work with sequences from very conserved regions (such as protease or RT), check [here](#) for more tips on identifying problem sequences.

1. Create a phylogenetic tree that includes all the sequences in the study. Common signs of trouble are:
 - Extreme intrapatient divergence
 - Extreme interpatient similarity
 - Mixed clusters (sequences from patient A clustering with patient B)

A phylogenetic tree clarifies the relations between the sequences. If you have lab strain contamination or sample mix-ups between two patients, a phylogenetic tree will likely show it. Once you have your sequences aligned, generate a simple neighbor joining tree to check for potential problems.

[Make Tree](#)
2. Compare your sequences to all published sequences (BLAST search).
 - Watch for lab strains with high similarity scores
 - Keep in mind that 100% identity is not required for contamination! ([example](#))

[BLAST](#)

Blast is a program that finds sequences with very high similarity to the query sequence. If your sequence is very similar to a published strain, especially a lab strain that is used for in vitro studies, it is likely that you have contamination. Even if your sequence is not identical to the lab strain, watch out for in vitro recombination, where only part of the sequence matches the lab strain, and the other part is derived from your patient sample. You can compare your sequences to all Genbank entries ([Genbank-BLAST](#)), which contains the very latest sequences, or against the HIV database (click the BLAST button) which can lag behind a few weeks, but contains more background information about the sequences.

What is 'reasonable similarity' depends on the gene or region (RT sequences are much more similar than V3 sequences) and on the population (compare a set of clonal sequences from different tissues of one person to a set from different persons in a clustered outbreak, to a set from different African countries). We've prepared some basic [guidelines](#).
3. Look carefully at the alignments, and pay attention to patient signature patterns.

Signature patterns often help to show what is 'typical' and 'atypical' for a patient, and thus help to recognize sequences that don't seem to belong with a patient. The usefulness of signature patterns can be seen in the contamination [examples](#). You can use [Vespa](#) to find the patterns, but often a simple alignment is sufficient to spot suspicious sequences. When you have an alignment, you can use [SeqPublish](#) to create a formatted version of it that will

N-Glycosite

- Tracks of patterns of N-linked glycosylation site (N-X-[ST]) change in sequences
- INPUT: A sequence alignment of interest
- OUTPUT:
 - Tallies of numbers of Ngly sites in each sequence
 - Highlighted Ngly sites
 - Graphics illustrating frequency of Ngly patterns in the alignment, and in sub-regions of the alignment
 - Frequencies of different patterns of X and Y in N-X-[ST]-Y

HIV Sequence Database: GLYCOSITE Submission Form - Microsoft Internet Explorer

Address <http://www.hiv.lanl.gov/content/hiv-db/GLYCOSITE/glycosite.html>

N-GLYCOSITE

This tool highlights and tallies N-linked glycosylation sites* in an aligned set of protein sequences.
(If you just want to tally the number of N-glycosylation sites, protein sequences do not need to be aligned.)

Submit your alignment in [FASTA](#) format here: [Sample Input](#)

* Note:

1. The most widely used N-linked glycosylation site pattern, NX[ST] (where X can be any amino acid), is called a sequon. This pattern forms a basis for most of the analyses on this web site. The extent of N-linked glycosylation of a particular N-linked glycosylation site, however, can be influenced by the content in which it is embedded, and could be expanded to a four amino acid NX[ST]Y patten, where the amino acid in the X or Y position of NX[ST]Y pattern can be important determinants of N-linked glycosylation efficiency. A particular strong effect is that a proline in position X or Y does not favor N-linked glycosylation. Thus we provide NX[ST] or NX[ST]Y summaries.
2. O-linked glycosylation signals are more difficult predict in protein sequences than N-linked sites, but one can estimate their positions in sequences using the [NetPhos](#) program.

References:
1) Marshall RD, Biochem Soc Symp. 40:17-26 (1974)
2) Kasturi et al., Biochem J. 323 (Pt 2):415-9 (1997)
3) Mellquist JL et al., Biochemistry. 37(19):6833-7 (1998)

Questions or comments? Contact us at seq-info@t10.lanl.gov

Last modified: Tue Feb 10 13:50 2004

[HCV Databases](#)

[Disclaimer/Privacy](#)

Immunology Database Overview

- HIV T-Cell (CTL, T-helper) and Antibody (Ab)
- Types of data recorded
 - ☐ Epitope sequence and location: HXB2 numbering, subtype
 - ☐ Immunogen
 - ☐ Host HLA or MHC, and Ab isotype
 - ☐ Notes summarize main findings
- Contents: data from 1985 through 2003
- Data from 2004 is being added and will be available early next year
 - ☐ 2618 CTL entries
 - ☐ 726 T-helper entries
 - ☐ 1223 Ab entries

The screenshot shows the HIV Molecular Immunology Database website in a Microsoft Internet Explorer browser window. The address bar displays the URL: <http://www.hiv.lanl.gov/content/immunology/index>. The website layout includes a left sidebar with navigation links such as "Immunology DB", "DB Help", "CTL search", "T Helper search", "AB search", "Epitope Maps", "HLA/TEM", "Tools & Links", "Home", "Immunology Tools", "Epitope", "PeptGen", "Motif Scan", "Sequence Locator", "ELF", "Publications", "FAQ", "Alignments", "Tutorials", "Reviews", "Compendia", "Links", and "Databases". The main content area is titled "HIV Molecular Immunology" and features a "News" section with items since 18 November 2004, a "Search the HIV Molecular Immunology Database" section with links to "Database Help", "CTL Search", "T Helper Search", and "Antibody Search", a "Database Products" section with links to "About the database", "Epitope maps for all proteins", "Epitope summary tables", "CTL epitopes", "T helper epitopes", "Antibody epitopes", "Antibody index by name", "Antibody index by binding type", "The HIV immunology compendia in PDF format", "Reviews from the 2002 and earlier Compendia in HTML and PDF formats", and "How to cite this database", an "Also available" section with links to "HLA Typing and Epitope Mapping", "Tools for immunologists", "Epitope Align", "PeptGen", "Motif Scan", "HIV/SIV Sequence Locator Tool", "ELF Epitope Location Finder tool", "Subtype and M group consensus and ancestral sequences", "Sequence tools", and "Other immunology tools and links", and a "Background" section. The footer of the website lists the compendium editors: Bette Korber, Christian Brander, Barton F. Haynes, Richard Koup, John P. Moore, Bruce D. Walker, and David I. Watkins, and the publisher: Los Alamos National Laboratory: Theoretical.

Immunology Database: Search

■ T Cells

- ☐ Cytotoxic T Lymphocytes (CTL)
- ☐ Helper T Lymphocytes (T-helper)
- ☐ Biological distinction between CTL and T-helper is not always obvious
- ☐ Organization is identical for CTL and T-helper
- ☐ One reference per entry

■ B Cells (Antibodies)

- ☐ One entry for each monoclonal antibody
- ☐ Many references per entry (up to 150)

HIV Immunology CTL, CD8+ T-Cell, Search

Proteins with defined epitopes: - ALL - p17 p17-p24 p24 p24-p2p7p1p6

Proteins with undefined epitopes: - ALL - Gag Gag/Pol Pol Vif

Epitope: - ALL - computer prediction HIV-1 and HCV co-infection HIV-1 exposed seronegative HIV-1 infected monocyte-derived HIV-1 infection HIV-1 or HIV-2 infection

Immunogen: - ALL - computer prediction HIV-1 and HCV co-infection HIV-1 exposed seronegative HIV-1 infected monocyte-derived HIV-1 infection HIV-1 or HIV-2 infection

If Immunogen is Vaccine, additional search details

Vaccine details: Vaccine type: - ALL - Vaccine strain: - ALL - Vaccine component: - ALL - Adjuvant: - ALL -

Species: - ALL -

HLA: - ALL - - NULL - A*0101 A*02 A*0201 A*0201 and Cw*08 A*0201, B*3501

Author:

Keywords: - ALL -

Search Reset

Los Alamos NATIONAL LABORATORY

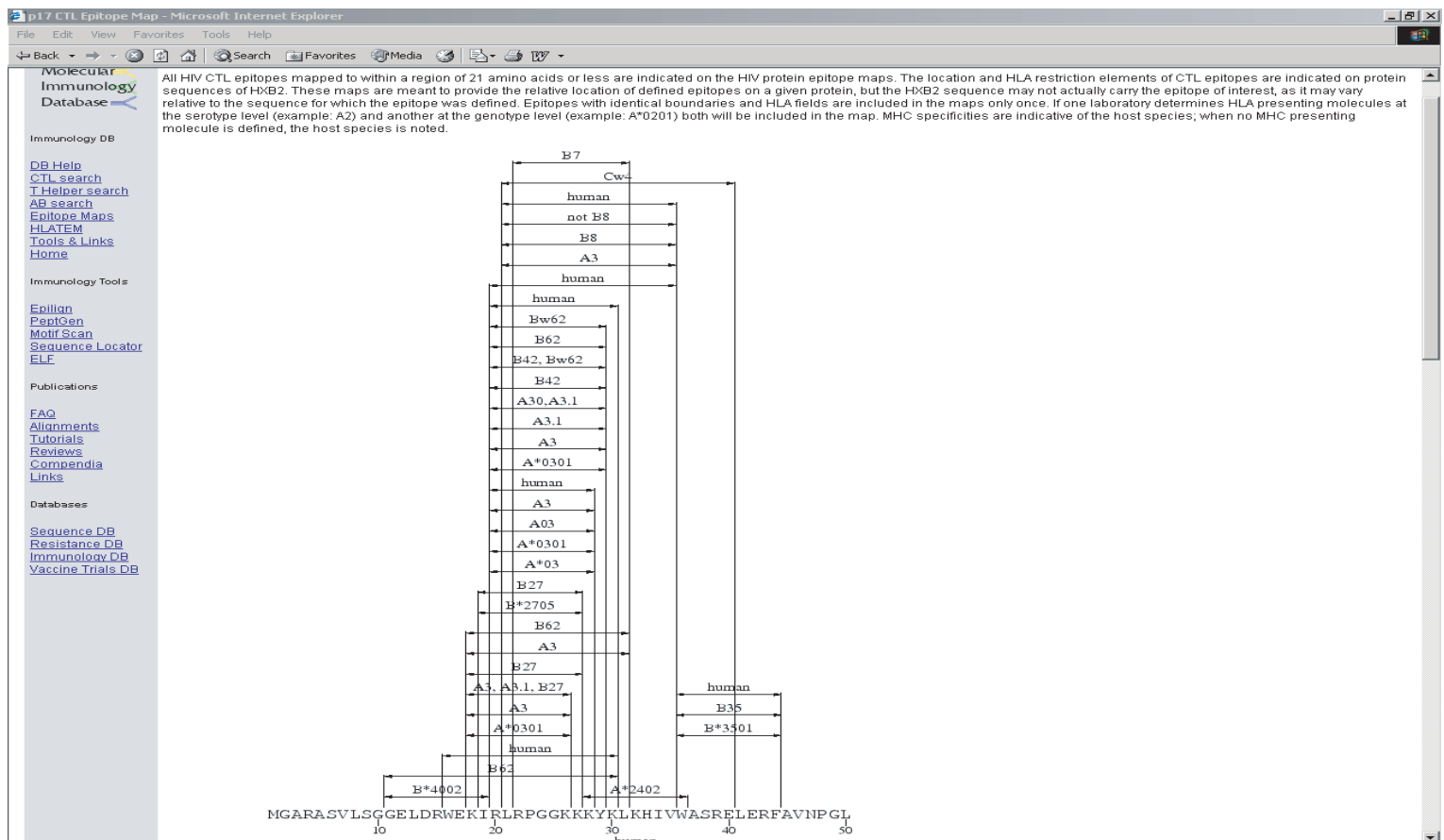
Operated by the University of California for the US Department of Energy Contract #AC02-84OR21400

NATIONAL INSTITUTES OF HEALTH

INTERNET

Immunology DB: Additional Information

- All entries for a reference
- Medline links to papers
- Epitope Tables
- Epitope Maps
 - Unique species/HLA for T cell epitopes
 - MAb name, species code for Ab
- Epitope Alignments
 - Extracted from HIV-sequence database, includes subtype, country and year of sampling



EPILIGN

- Generates an alignment of your HIV-1 amino acid sequence against our web alignments
- Can be used to align epitopes, functional domains, or any protein region of interest
- Sequence names include subtype, country and year of sampling

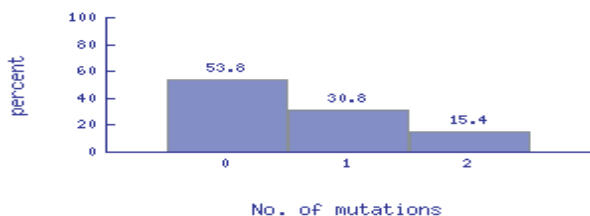
Example of output:

Summary for subtype A1

Variant	Count	Percent
SLYNTVATL	7	53.8
--F-----	3	23.1
--F----V-	2	15.4
-----V-	1	7.7

Total sequences = 13
Number of variants = 4

Mutation percentages



Query:	SLYNTVATL
Query Length:	9
HXB2 Location:	Gag 77-85 = p17 77-85
Alignment:	GAG, 458 sequences

Summarize

Query	SLYNTVATL
A1.KE.86.ML170	--F-----
A1.KE.94.Q23	--F-----
A1.SE.94.SE7253	--F----V-
A1.SE.94.SE7535	-----V-
A1.SE.95.SE8538	-----V-
A1.SE.95.SE8891	-----V-
A1.SE.95.UGSE8131	-----V-
A1.TZ.97.97T203	--F----V-

HLA Binding Motif Scanner : MotifScan

- Finds HLA anchor motifs within protein sequences for specified HLA genotypes, serotypes, or supertypes
- HLA anchor motif dictionaries are available on line
- Main motif and supermotif sources:
 - ☐ SYPHEITHI Database, Rammensee *et al.* www.syfpeithi.de
 - ☐ HLA Facts Book, Marsh *et al.* 2000
 - ☐ Sette & Sidney, *Immunogenetics* **50**:201-212, 1999
 - ☐ Isobella Honeyborne and Philip Goulder, Yusim *et al* 2003, (HLA-C motifs predicted by genetic similarities in HLA molecules)
- **INPUT:**
 - ☐ User defined query sequence or aligned sequences, or reference set
 - ☐ Selected HLA anchor motifs, or user defined motif
 - ☐ The user defined motif function could be used to search for other patterns of interest in sequences
- **OUTPUT:**
 - ☐ Anchor residue positions are highlighted in the query sequence
 - ☐ Potential epitopes and positions are listed
 - ☐ Output can be downloaded as text, convenient for further analysis
- **NEW FEATURE: Finds all known motifs in a sequence**

HLA Binding Motif Scanner

Use this page to find HLA anchor residue motifs within protein sequences for specified HLA genotype, serotypes or supertypes. Refer to the [Help](#) page for more information.

Please enter your search criteria

Genotype	Serotype	Supertype
A*01	A1	A1
A*0101	A2	A2
A*0201	A3	A3
A*0202	A11	A24
A*0204	A24(9)	B7
A*0205	A25(10)	B27
A*0206	A26(10)	B44

Motif Source ☒ Marsh2000 ☒ SYFPEITHI ☒ Others

Motif Length ☐ 8 ☒ 9 ☐ 10 ☐ 11

Custom Motif

Search **Reset**

Find All Motifs in a Sequence

Enter a protein sequence below to scan it for all known motifs.

Scan All **Reset**

Data dictionaries

[View](#) or [download](#) the HLA genotype/serotype dictionary.

[View](#) or [download](#) the HLA genotype/motif dictionary.

[View](#) or [download](#) the HLA supertype dictionary.

HLA Binding Motif Scanner - Microsoft Internet Explorer

File Edit View Favorites Tools Links Address http://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan?genotype=A*0202&source=Marsh2000&source=SYFPEITHI&source=Others&length=9&custom_motif=&search= Go

Home

Immunology Tools

Epilign
PeptGen
Motif Scan
Sequence Locator
ELF

Publications

FAQ
Alignments
Tutorials
Reviews
Compendia
Links

Databases

Sequence DB
Resistance DB
Immunology DB
Vaccine Trials DB

Your search results

Genotype	Serotype	Motif	Source	Scan?
A*0202	A2	x-[L]-x-x-x-x-x-x-[L]	Marsh2000	<input checked="" type="checkbox"/>
A*0202	A2	x-[L(A)]-x-x-x-x-x-x-[LV]	SYFPEITHI	<input checked="" type="checkbox"/>

Sequence selection

Please select the HIV sequences you wish to scan for motifs.

Predefined sequences:

☒ HXB2 ☐ Vif consensus ☐ Tat consensus ☐ Nef consensus
☐ Gag consensus ☐ Vpu consensus ☐ Vpr consensus ☐ Vpx consensus
☐ Pol consensus ☐ Rev consensus ☐ Env consensus

The consensus sequences are sets of [Consensus and Ancestral Sequences for M and O Groups](#) from the [LANL HIV Sequence Database](#).

Input protein sequences in FASTA or TABLE format:

Submit a file of protein sequences in FASTA or TABLE format:

Are the input sequences aligned?

☐ Yes ☒ No

HLA Binding Motif Scanner - Microsoft Internet Explorer

File Edit View Favorites Tools Links Address http://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan Go

Genotype	Serotype	Supertype	Motif	Source
A*0202	A2		x-[L]-x-x-x-x-x-x-[L]	Marsh2000

Matches (Download as FASTA)

	10	20	30	40	50
Test.5B	ktiifkpsag	gdpeivthaf	ncggeffycn	ttklfnstwn	stwd----ln
Test.6B	ktiafkssg	gdpeivthaf	ncggeffhcn	stqlfnstwn	gnd-----td
Test.7B	ktivfnsgg	gdpeivthaf	ncggeffycn	taqlfnstwn	-----
Test.5B	qtanh-egnd	--titlpcrl	kqivnm-wqe	vgkamyappi	eggiacfshi
Test.6B	tketn-dtag	--titlpcrl	kqivnl-wqe	vgkamyappi	rgqircssni
Test.7B	--nt-iens	--titlpcrl	kqivnm-wqe	vgkamyappi	rgqircssni
Test.5B	tgliltddgg	-nd-----	tnnetfrpqq	gnmkdnwrse	lykykvvkik
Test.6B	tgliltddgg	-ni--t---	netelfrpgg	gdmdnwrse	lykykvvkik
Test.7B	tgliltddgg	-nng-s---	nttetfrpqq	gnmdnwrse	lykykvvkik
Test.5B	plgiaptkak	rrvvqrekra	v-gtlgamfl	g-flgaagst	mgaasvtl
Test.6B	plgvaptkak	rrvvqrekra	v-gtlgamfl	g-flgaagst	mgaasvtl
Test.7B	plgvaptkak	rrvvqrekra	v-gtlgamfl	g-flgaagst	mgaasitl

List of potential epitopes (Download as TSV)

Protein Seq.	Pos.	Aln. Pos.	Sequence Anchors
Test.5B	88-96	96-104	CLSHITGLL.L.....L

References

Marsh2000 Steven G. E. Marsh, Peter Parham, and Linda D. Barber. *The HLA FactsBook*. Academic Press, San Diego, 2000.

SYFPEITHI The SYFPEITHI Database of MHC Ligands, Peptide Motifs and Epitope Prediction. Jan 2003. URL: <http://syfpeithi.bmi-heidelberg.com/>

HLA Binding Motif Scanner - Microsoft Internet Explorer

File Edit View Favorites Tools Links Address http://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan?seq_input=MENRWQVMIVWQVDRMRIRTWKSLVKHHMYVSGKARGWFYRHYESPHRISSEVHIPGLDARLVITYWGLH Go

HIV Molecular Immunology Database

Immunology DB

DB Help
CTL search
T Helper search
AB search
Epitope Maps
HLAEM
Tools & Links
Home

Immunology Tools

Epilign
PeptGen
Motif Scan
Sequence Locator
ELF

Publications

FAQ
Alignments
Tutorials
Reviews
Compendia
Links

Databases

HLA Binding Motif Scanner

Scanning
MENRWQVMIVWQVDRMRIRTWKSLVKHHMYVSGKARGWFYRHYESPHRISSEVHIPGLDARLVITYWGLHTGERDWHLGQGVSEIWRKKRYSTQVDPPELADQLII for all motifs.

Matches (Download as FASTA)

```
>A*01 x-x-[DE]-x-x-x-x-x-[Y] [SYFPEITHI]
menrwqvmiv wqvdrmrirt wkslvkhhmy vsgkargwfy rhhyesphpr 50
issevhiplg darlvittyw glhtgerdwh lqggvsiewr kkrystqvdp 100
elaqlinh y yfcdcfdsai rkallghivs prceyqaghn kvgslylqal 150
aalitpkkik pplsavtklt edrwnkpqkt kggrshtmn gh 192

>A*0101 x-x-[DE]-x-x-x-x-x-[Y] [Marsh2000]
menrwqvmiv wqvdrmrirt wkslvkhhmy vsgkargwfy rhhyesphpr 50
issevhiplg darlvittyw glhtgerdwh lqggvsiewr kkrystqvdp 100
elaqlinh y yfcdcfdsai rkallghivs prceyqaghn kvgslylqal 150
aalitpkkik pplsavtklt edrwnkpqkt kggrshtmn gh 192

>A*0201 x-[LM]-x-x-x-x-x-x-[VL] [SYFPEITHI]
menrwqvmiv wqvdrmrirt wkslvkhhmy vsgkargwfy rhhyesphpr 50
issevhiplg darlvittyw glhtgerdwh lqggvsiewr kkrystqvdp 100
elaqlinh y yfcdcfdsai rkallghivs prceyqaghn kvgslylqal 150
aalitpkkik pplsavtklt edrwnkpqkt kggrshtmn gh 192

>A*0201 x-[L(M)]-x-x-x-x-x-x-[V(L)] [Marsh2000]
menrwqvmiv wqvdrmrirt wkslvkhhmy vsgkargwfy rhhyesphpr 50
issevhiplg darlvittyw glhtgerdwh lqggvsiewr kkrystqvdp 100
elaqlinh y yfcdcfdsai rkallghivs prceyqaghn kvgslylqal 150
aalitpkkik pplsavtklt edrwnkpqkt kggrshtmn gh 192

>A*0202 x-[L]-x-x-x-x-x-x-[L] [Marsh2000]
menrwqvmiv wqvdrmrirt wkslvkhhmy vsgkargwfy rhhyesphpr 50
issevhiplg darlvittyw glhtgerdwh lqggvsiewr kkrystqvdp 100
elaqlinh y yfcdcfdsai rkallghivs prceyqaghn kvgslylqal 150
aalitpkkik pplsavtklt edrwnkpqkt kggrshtmn gh 192

>A*0202 x-[L(A)]-x-x-x-x-x-x-[LV] [SYFPEITHI]
menrwqvmiv wqvdrmrirt wkslvkhhmy vsgkargwfy rhhyesphpr 50
issevhiplg darlvittyw glhtgerdwh lqggvsiewr kkrystqvdp 100
elaqlinh y yfcdcfdsai rkallghivs prceyqaghn kvgslylqal 150
aalitpkkik pplsavtklt edrwnkpqkt kggrshtmn gh 192
```

Hepitope: Identifying epitopes in reactive peptides

- Input1: A set of the HLA alleles for a study set of individuals
- Input2: A list of peptides and the individuals that react with each peptide
- Output: Listings of HLA alleles that are found in people that react with the peptide, ordered by those that are enriched among reactive people

HLA Enriched Epitope Possible

Introduction

The Hepitope tests for HLA alleles that are enriched in individuals that react with a set of peptides. This can be used in conjunction with our ELF program, which will scan a peptide for known epitopes in the database and anchor motifs to help identify epitopes within a larger peptide fragment (Hopeful Epitopes, or Hepitopes). See below for details about the input and output of the program.

Input

Using Sample Input:

List of patient HLA alleles:

Patient1	A*0201	A*0201	B*5703	B*1701	Cw*0701	Cw*0705
Patient2	A*0201	A*0701	B*1202	B*0801	Cw*0701	Cw*0401
Patient3	A*1101	A*2403	B*0801	B*5801	Cw*0701	Cw*1501
Patient4	A*3002	A*3002	B*5802	B*5802	Cw*0602	Cw*0602

List of reactive peptides:

MGARASVLSGGELDRWEK	Patient1
SGGELDRWEKIRLRPGGK	Patient2
EKIRLRPGGKKYKIKHI	Patient3
	Patient4

Results

Peptide	HLA Type	a	b	c	d	P
MGARASVLSGGELDRWEK	B*1701	1	0	0	3	0.25000000
	B*5703	1	0	0	3	0.25000000
	Cw*0705	1	0	0	3	0.25000000
	A*0201	1	0	1	2	0.50000000
	Cw*0701	1	0	2	1	0.75000000
Patient HLA						
Patient1 A*0201 A*0201 B*1701 B*5703 Cw*0701 Cw*0705						
SGGELDRWEKIRLRPGGK	B*0801	2	0	0	2	0.16666667
	A*0701	1	1	0	2	0.50000000
	A*1101	1	1	0	2	0.50000000
	A*2403	1	1	0	2	0.50000000
	B*1202	1	1	0	2	0.50000000
	B*5801	1	1	0	2	0.50000000
	Cw*0401	1	1	0	2	0.50000000
	Cw*0701	2	0	1	1	0.50000000
	Cw*1501	1	1	0	2	0.50000000
	A*0201	1	1	1	1	0.83333333
Patient HLA						
Patient2 A*0201 A*0701 B*0801 B*1202 Cw*0401 Cw*0701						
Patient3 A*1101 A*2403 B*0801 B*5801 Cw*0701 Cw*1501						
EKIRLRPGGKKYKIKHI	A*3002	1	0	0	3	0.25000000
	B*5802	1	0	0	3	0.25000000
	Cw*0602	1	0	0	3	0.25000000
	Patient HLA					
Patient4 A*3002 A*3002 B*5802 B*5802 Cw*0602 Cw*0602						

ELF – Epitope Location Finder

- Input1: A set of HLA alleles of interest
- Input2: A reactive peptide
- Output: Listings of HLA alleles that are possible based on HLA motifs, and listings of all known epitope that are located within the peptide.

http://hiv.lanl.gov/content/hiv-db/ELF/epitope_analyzer.html?sample_input=1 - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://hiv.lanl.gov/content/hiv-db/ELF/epitope_analyzer.html?sample_input=1

ELF
Epitope Location Finder

This is a beta-test version of the HIV Molecular Immunology Database "ELF" site. Please email comments, suggestions, and problems to ccalef@lanl.gov or btik@lanl.gov.

Purpose

ELF searches a submitted protein sequence for the occurrence of any epitopes known from our immunology database. Those epitopes whose HLA agrees with the submitted HLA(s) are flagged. Short peptides within the protein, "potential epitopes," whose sequence agrees with the binding motifs of the submitted HLAs are also displayed. Maps can be prepared that highlight every known epitope of the submitted HLA across the HIV proteome. [More detail.](#)

How to use

In the HLA box enter one or more HLA types. If you leave the box blank or type "all" or * in the box, all HLAs will be examined. Then paste your HIV protein sequence (in raw format, i.e. only the amino acids themselves) into the text box. There are two checkboxes at the bottom. Show all epitopes, the default, will find all epitopes in our database within the bounds of your submitted protein regardless of their HLA. If you uncheck this box the program will find only known database epitopes whose HLA agrees with your submitted HLAs. Checking the Show Maps box will cause the program to prepare maps that highlight every known epitope of the submitted HLA across the HIV proteome.

HLA A2, B44

Protein sequence PQITLWQRPVLVTIKIGGQLKEALLDTGADDTVLEDNNLPGRWPKPMIGGIGGFIVKVKQ

Show all known epitopes ☒ **Draw maps?** ☒ No ☐ Only this protein ☐ All proteins

Submit **Reset**

Last modified: Fri May 6 17:00 2005


Questions or comments? Contact us at seq-info@t10.lanl.gov

http://hiv.lanl.gov/cgi-bin/ELF/epitope_analyzer.cgi - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address http://hiv.lanl.gov/cgi-bin/ELF/epitope_analyzer.cgi

Epitopes from our database aligned to your query sequence

Bold letters colored **red** indicate residues in known epitopes which differ from the equivalent residues in the query sequence. The symbol  means the epitope's HLA matches one of your submitted HLAs.

Download this alignment in format:

table
fasta
mase
pretty

PQITLWQRPVLVTIKIGGQLKEALLDTGADDTVLEDNNLPGRWPKPMIGGIGGFIVKVKQ

ITLWQRPVLV A*6802, A*7401, A19 align

ITLWQRPVLV A*6802 align

ITLWQRPVLV A*7401 align

ITLWQRPVLV A28 align

ITLWQRPVLV A28 supertype align

ITLWQRPVLV A74 align

ITLWQRPVLV A2 align

ITLWQRPVLV A*3303 align

ITLWQRPVLV A3 supertype align

ITLWQRPVLV A*1101 align

ITLWQRPVLV A3 supertype align

ITLWQRPVLV A*6802 align

ITLWQRPVLV A*6802 align

ITLWQRPVLV A*6802 align

ITLWQRPVLV B*44 align

ITLWQRPVLV B44 align

ITLWQRPVLV A2 supertype align

Potential "epitopes" in your input sequence

These peptides have C-term anchor residues, highlighted in **blue**, and internal anchors highlighted in **magenta**. These anchor positions match one or more motifs associated with the submitted HLA, but are *not* found in our database.

Download this alignment in format:

table
fasta
mase
pretty

PQITLWQRPVLVTIKIGGQLKEALLDTGADDTVLEDNNLPGRWPKPMIGGIGGFIVKVKQ

PQITLWQRPVLV (A*0205 . [VLIMQ] [L])

PQITLWQRPVLV (A*0214 . [VQL] [LV])

PQITLWQRPVLV (A*0205 . [VLIMQ] [L])

TLWQRPVLV (A*0201 . [LM] [VL])

TLWQRPVLV (A*0202 . [L] [LV])

TLWQRPVLV (A*0214 . [VQL] [LV])

LVTIKIGGQL (A*0205 . [VLIMQ] [L])

LVTIKIGGQL (A*0214 . [VQL] [LV])

TIKIGGQL (A*0205 . [VLIMQ] [L])

KIGGQLKEALL (A*0205 . [VLIMQ] [L])

QKLEALL (A*0205 . [VLIMQ] [L])

QKLEALL (A*0214 . [VQL] [LV])

LDTGADDTV (A*0201 . [LM] [VL])

LDTGADDTV (A*0202 . [L] [LV])

LDTGADDTV (A*0214 . [VQL] [LV])

ITLWQRPVLV (A*0205 . [VLIMQ] [L])

TVLEDNNL (A*0214 . [VQL] [LV])